DEPARTMENT OF BIG DATA ANALYTICS

ST.JOSEPH'S COLLEGE (AUTONOMOUS)

BENGALURU - 560027

ADVANCED STATISTICS LAB
PROJECT REPORT

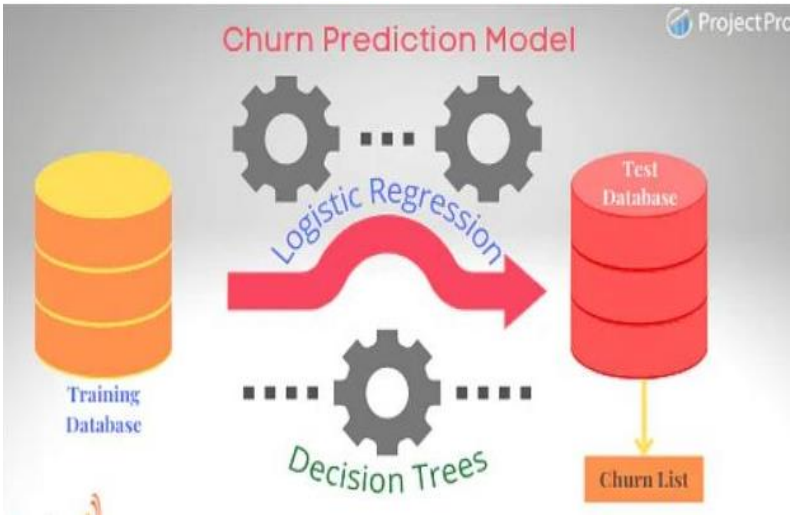
Submitted by:
Alexei Bueno.H - 21BDA01
Shruti Goyal - 21BDA51
Sreelakshmi cv - 21BDA39

CUSTOMER CHURN PREDICTION IN TELECOM INDUSTRY

I. INTRODUCTION

Customer churn is referred to as the inclination of a customer to leave a service provider. Customer churn prediction is the process of identifying those customers who could leave or switch from the current service provider company due to certain reasons. The major aim of the churn prediction model is to identify such customers so that the retention strategies could be targeted upon them and the company may flourish by maximizing its overall revenue.

Customer churn prediction has been raised as a notorious issue in many fields such as telecommunication, Credit Card, Internet Service Providers, Electronic Commerce, Retail Marketing, Newspaper publishing companies, Banking and financial services. But among all other fields, the average churn rate in the telecom industry is approx. 1.9% per month across the four major carriers, but could rise as high as 67% annually for prepaid services.



II. PROBLEM STATEMENT

The telecommunications industry is made up of cable companies, internet service providers, satellite companies and telephone companies. Telecommunications is defined as communicating over a distance. The industry's origin can be traced to postal courier services.

Customer churn in the telecom industry poses one of the most significant risks to loss of revenue. The average churn rate in the telecom industry is approx. 1.9% per month across the four major carriers, but could rise as high as 67% annually for prepaid services.

Considering that customer churn in telecom is expensive and inevitable, leveraging to understand the factors that influence customer attrition, identifying customers that are most likely to churn and offering them discounts can be a great way to reduce it.

In our project, we build some models to understand the correlation between the different variables in the dataset and customer churn. Then we check the accuracy of those modes to find which model is best. This end-to-end churn prediction model will tweak the problem of dissatisfaction and make the revenue flowing for the telecom company.

Keywords: Churn, Logistic Regression, Support Vector Machine, Decision Tree, PCA(principal component analysis)

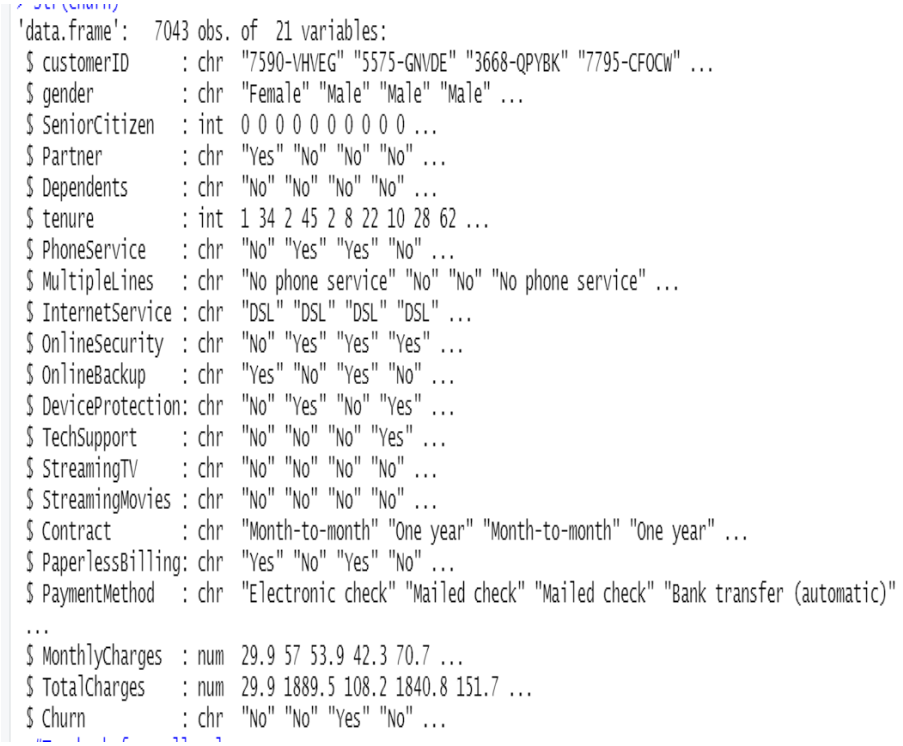
III. DATASET

i. ABOUT THE DATA

The data set that we used in our project was retrieved from the Kaggle website.

Source:https://www.kaggle.com/datasets/zagarsuren/telecom-churn-dataset-ibm-watson-analytics

ii. STRUCTURE OF THE DATASET



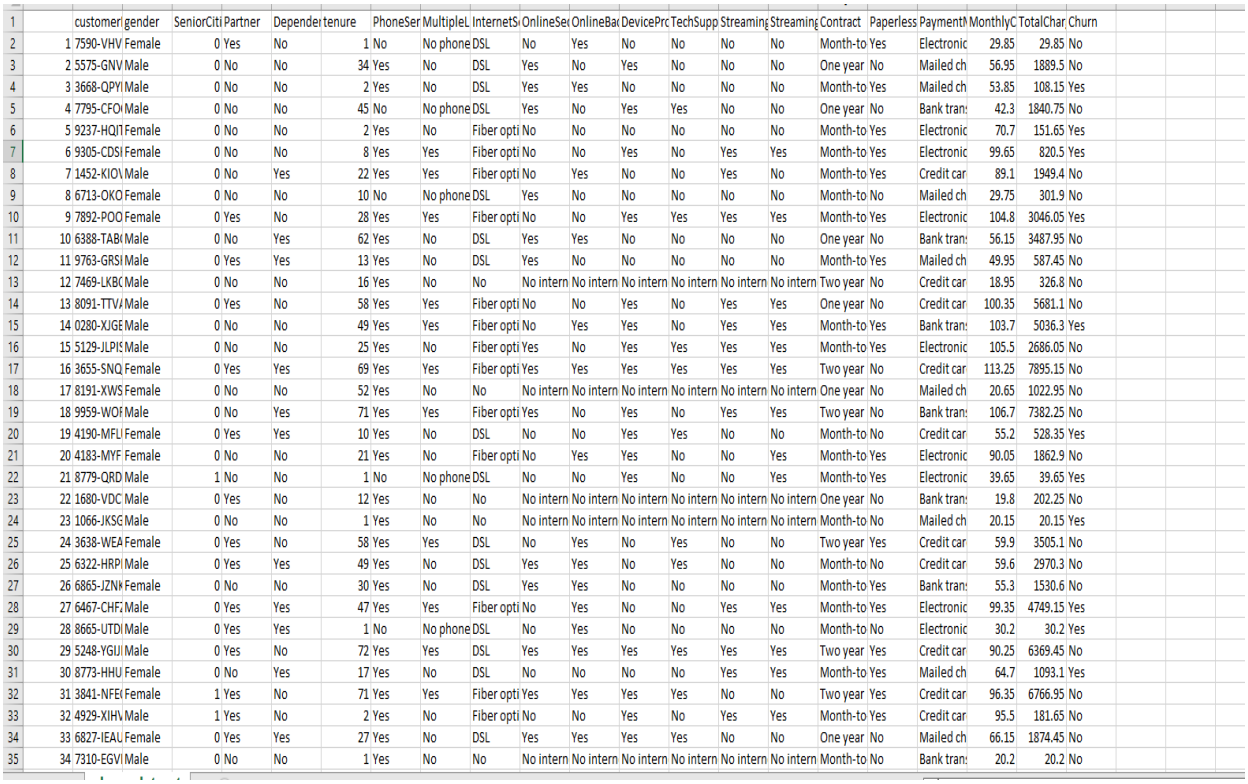
```
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr  "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr  "Yes" "No" "No" "No" ...
 $ Dependents      : chr  "No" "No" "No" "No" ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr  "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr  "No" "No" "No" "No" ...
 $ StreamingMovies : chr  "No" "No" "No" "No" ...
 $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)"
 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr  "No" "No" "Yes" "No" ...
```

iii. ATTRIBUTE INFORMATION

Our dataset contains variables like gender, partner, churn(that is whether the customer churn or not)etc., and our data contains 7043 rows and 21 columns. In our project the target variable is churn.

The dataset contains the information about :

- Customers who left within the last month - the column is called the churn.
- Services that each customer has signed up for - phone, multiple lines, internet, online security, online backup, device protection, tech support and streaming TV and movies.
- Customer account information - how long they have been a customer, contract, payment method, paperless billing, monthly charges and total charges
- Demographic information about customers - gender, age range and if they have partners and dependencies.



	customer	gender	SeniorCiti	Partner	Dependen	tenure	PhoneSer	MultipleL	InternetS	OnlineSe	OnlineBa	DevicePro	TechSupp	Streaming	Streaming	Contract	Paperless	Payment	MonthlyC	TotalChar	Churn
1	7590-VHV	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to	Yes	Electronic	29.85	29.85	No
2	5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed ch	56.95	1889.5	No
3	3668-QPY	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to	Yes	Mailed ch	53.85	108.15	Yes
4	7795-CFO	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank tran	42.3	1840.75	No
5	9237-HQI	Female	0	No	No	2	Yes	No	Fiber opti	No	No	No	No	No	No	Month-to	Yes	Electronic	70.7	151.65	Yes
6	9305-CDS	Female	0	No	No	8	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	Month-to	Yes	Electronic	99.65	820.5	Yes
7	1452-KIO	Male	0	No	Yes	22	Yes	Yes	Fiber opti	No	Yes	No	No	Yes	No	Month-to	Yes	Credit car	89.1	1949.4	No
8	6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to	No	Mailed ch	29.75	301.9	No
9	7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber opti	No	No	Yes	Yes	Yes	Yes	Month-to	Yes	Electronic	104.8	3046.05	Yes
10	6388-TAB	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank tran	56.15	3487.95	No
11	9763-GRS	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to	Yes	Mailed ch	49.95	587.45	No
12	7469-LKB	Male	0	No	No	16	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	Two year	No	Credit car	18.95	326.8	No
13	8091-TTV	Male	0	Yes	No	58	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	One year	No	Credit car	100.35	5681.1	No
14	0280-XJG	Male	0	No	No	49	Yes	Yes	Fiber opti	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	Bank tran	103.7	5036.3	Yes
15	5129-JLP	Male	0	No	No	25	Yes	No	Fiber opti	Yes	No	Yes	Yes	Yes	Yes	Month-to	Yes	Electronic	105.5	2686.05	No
16	3655-SNQ	Female	0	Yes	Yes	69	Yes	Yes	Fiber opti	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit car	113.25	7895.15	No
17	8191-XWS	Female	0	No	No	52	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	One year	No	Mailed ch	20.65	1022.95	No
18	9959-WOF	Male	0	No	Yes	71	Yes	Yes	Fiber opti	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank tran	106.7	7382.25	No
19	4190-MFL	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to	No	Credit car	55.2	528.35	Yes
20	4183-MYF	Female	0	No	No	21	Yes	No	Fiber opti	No	Yes	Yes	No	No	Yes	Month-to	Yes	Electronic	90.05	1862.9	No
21	8779-QRD	Male	1	No	No	1	No	No phone	DSL	No	No	Yes	No	No	Yes	Month-to	Yes	Electronic	39.65	39.65	Yes
22	1680-VDC	Male	0	Yes	No	12	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	One year	No	Bank tran	19.8	202.25	No
23	1066-JKS	Male	0	No	No	1	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	Month-to	No	Mailed ch	20.15	20.15	Yes
24	3638-WEA	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit car	59.9	3505.1	No
25	6322-HRP	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No	Month-to	No	Credit car	59.6	2970.3	No
26	6865-JZN	Female	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to	Yes	Bank tran	55.3	1530.6	No
27	6467-CHF	Male	0	Yes	Yes	47	Yes	Yes	Fiber opti	No	Yes	No	No	Yes	Yes	Month-to	Yes	Electronic	99.35	4749.15	Yes
28	8665-UTD	Male	0	Yes	Yes	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to	No	Electronic	30.2	30.2	Yes
29	5248-YGL	Male	0	Yes	No	72	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes	Yes	Two year	Yes	Credit car	90.25	6369.45	No
30	8773-HHU	Female	0	No	Yes	17	Yes	No	DSL	No	No	No	No	Yes	Yes	Month-to	Yes	Mailed ch	64.7	1093.1	Yes
31	3841-NFE	Female	1	Yes	No	71	Yes	Yes	Fiber opti	Yes	Yes	Yes	Yes	No	No	Two year	Yes	Credit car	96.35	6766.95	No
32	4929-XIH	Male	1	Yes	No	2	Yes	No	Fiber opti	No	No	Yes	No	Yes	Yes	Month-to	Yes	Credit car	95.5	181.65	No
33	6827-IEA	Female	0	Yes	Yes	27	Yes	No	DSL	Yes	Yes	Yes	Yes	No	No	One year	No	Mailed ch	66.15	1874.45	No
34	7310-EGV	Male	0	No	No	1	Yes	No	No	No intern	No intern	No intern	No intern	No intern	No intern	Month-to	No	Bank tran	20.2	20.2	No

IV. METHODOLOGY

Data preprocessing includes cleaning and integration of the data. We cleaned the data by removing the null values and changing values of some columns.

We use ggplot to plot the dataset. Ggplot is a plotting package that provides useful commands to create complex plots from data in a data frame.
We built a logistic regression(Logistic regression in R programming is a classification algorithm used to find the probability of event success and event failure), machine learning model to understand the correlation between the different variables in the dataset and to predict the customer churn count. We also test the accuracy.

We ran several tests for chi-square analysis and also tested the fitness of the model. Chi-square test in R is a statistical method which is used to determine if two categorical variables have a significant correlation between them.

V. DATA CLEANING

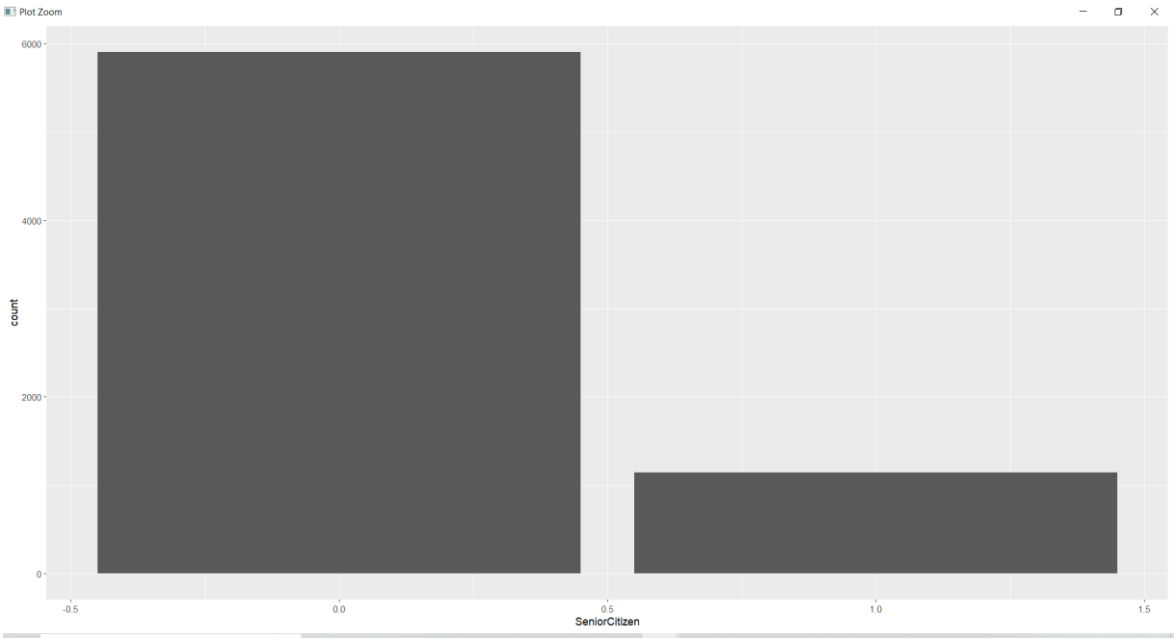
Data cleaning is the process to remove incorrect data, incomplete data and inaccurate data from the datasets, and it also replaces the missing values.

- For this, the first step is to find whether there are any missing values or not. We had 11 null values in the column Total Charges. So we removed those null values.
- Converted “No internet service” to “No” for one column : “Streaming Movies”
- Changed “No phone service” to “No” for “MultipleLines”
    - Minimum churn tenure is 1 month and maximum churn tenure is 72, so grouped them in different groups – a. 0-12 Month b.12-24 Month c. 24-48 Month d. 48-60 Month e. Greater than 60.
    - Changed the values of “Senior Citizen” from 0 or 1 to “No” or “Yes”

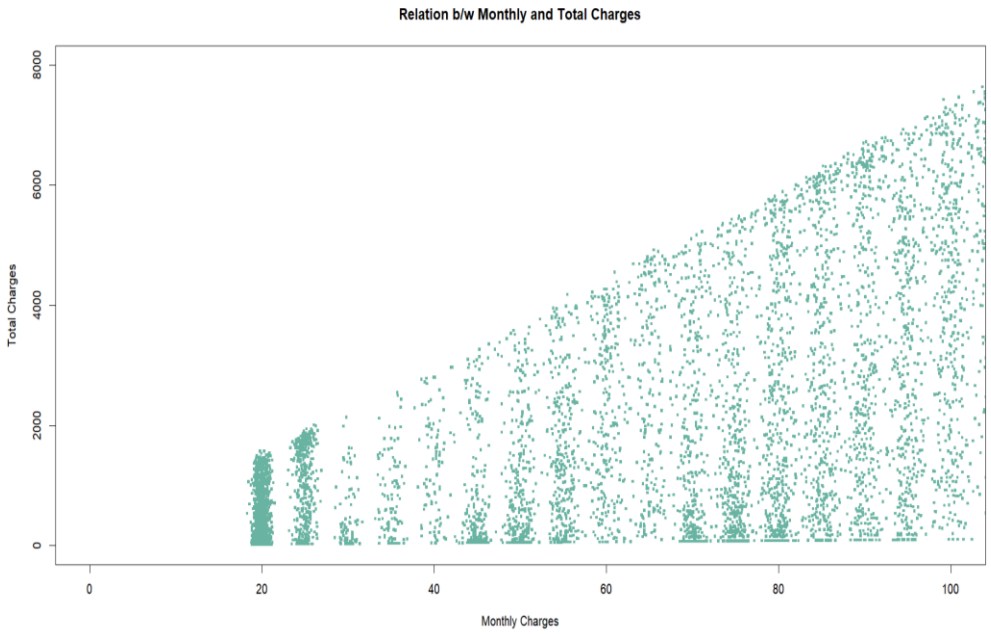
VI. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques.

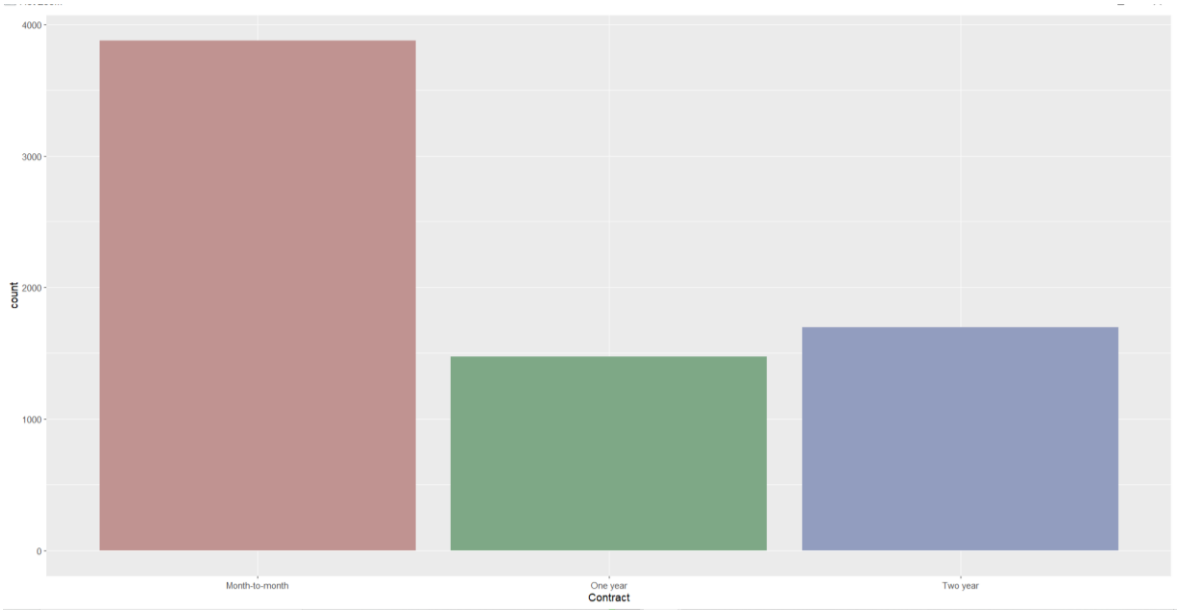
- In our dataset there are few senior citizens who are churning.



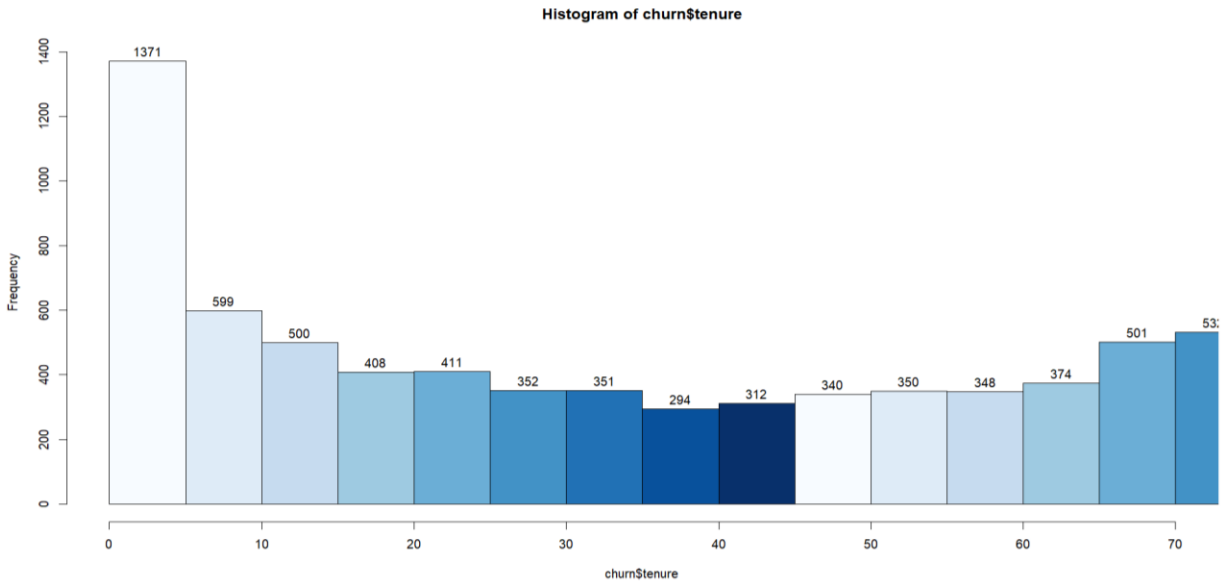
- When the total monthly charge of a customer is high(the total charges increases as the monthly bill for a customer increases) then the customers do not churn the telecom industry because when the customer switch to other companies the total charge is the same and that will not affect the customer so the customer need not change to other company.



- People prefer month-to-month contracts over yearly. Because they can easily change the plan or service provider. If it's a month, people prefer month-to-month contracts over yearly. Because they can easily change the plan or service provider if it's monthly.



- By plotting the graph, we can see that customers have been with the telecom company for just a month to 6 months, while quite a few are there for about 72 months.



VII. MODEL BUILDING

To check the dependencies of the categorical variables with the target variable we use the chi-square test of independence. If the p value is greater than 0.05 we reject the null hypothesis. Since the p-value for Gender, Phone Service is greater than 0.05, we reject the null hypothesis, hence these variables are not dependent on Churn(the target variable). Hence, we keep these two variables, and discard the rest.

Then we created a correlation matrix to check dependence of continuous variables and we found the Total charges are highly correlated with other variables. The only variables that are not correlated with the others are, Gender, Phone Service, Tenure and Monthly Charges.

VIII. CLASSIFICATION MODELS

Different machine learning models are suitable for various situations. In supervised learning, classification and regression are the two main categories of machine learning problems.
As we were dealing with the classification problem so the machine learning models that we used were:

1) Logistic Regression
2) Support vector machine
3) Decision Tree
4) Principal Component Analysis - To determine principal components.

1. LOGISTIC REGRESSION

Logistic regression is a supervised learning algorithm used to predict a dependent categorical target variable. In essence, if you have a large set of data that you want to categorize, logistic regression may be able to help. We fit logistic regression model t and then we calculate VIF values for each predictor in our model and the VIF values are below 5, which indicates that there is no multicollinearity.

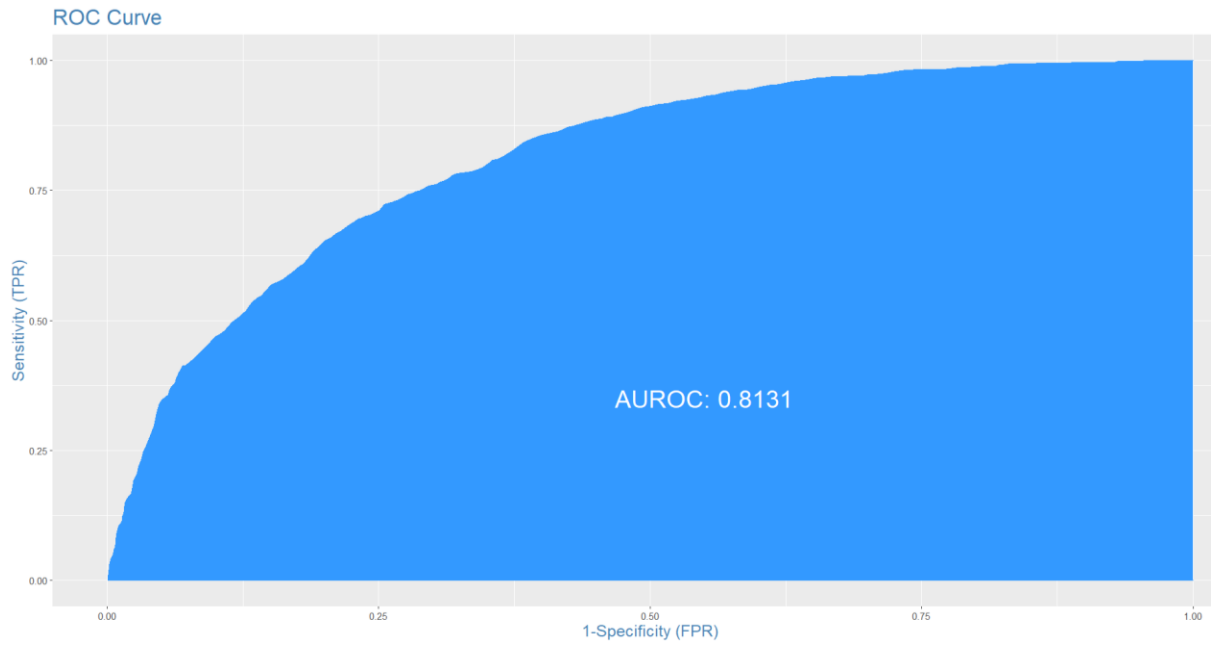
We also look at the confusion matrix(A confusion matrix is a performance measurement tool, often used for machine learning classification tasks where the output of the model could be 2 or more classes (i.e. binary classification and multiclass classification). The confusion matrix is especially useful when measuring recall, precision, specificity, accuracy, and the AUC of a classification model) and the below is our confusion matrix.

```
     0    1
0 4998 1400
1  176  469
```

To know the best threshold value for our logistic model, we calculated the f1 score, sensitivity and specificity for all the threshold values from 0 to 1. Then we exported the values into a csv file. The maximum f1 score that we got was 0.869542 which was at 0.591.

S.No.	auc	sensitity	f1Scores	specificity	cutoff
592	0.64678231	0.34082397	0.86954216	0.952740655	0.591
593	0.64634411	0.339753879	0.86948838	0.95293434	0.592
594	0.64529703	0.337078652	0.869403974	0.953515398	0.593
591	0.64675931	0.341359016	0.869319187	0.952159597	0.59
590	0.6471745	0.342964152	0.869149783	0.951384854	0.589

We did model diagnostics to evaluate how well our model performs on the test dataset.



2. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well, it's best suited for classification.

Steps:

1. Train the model and split.

2. We fit in linear and radial svm.

3. For both the models, we found the summary of the model by using “svmfit”.

4. At last we found the accuracy of this model for linear and radial svm by using a confusion matrix.

We find that the accuracy of both the svm model is approx. 0.765.

And the screenshot for one of the statistics output is as follows:

```
Confusion Matrix and Statistics

          0    1
  0 1390  159
  1  292  269

               Accuracy : 0.7863
                 95% CI : (0.7681, 0.8036)
    No Information Rate : 0.7972
    P-Value [Acc > NIR] : 0.8978

                  Kappa : 0.4077

 Mcnemar's Test P-Value : 5.112e-10

            Sensitivity : 0.8264
            Specificity : 0.6285
         Pos Pred Value : 0.8974
         Neg Pred Value : 0.4795
             Prevalence : 0.7972
         Detection Rate : 0.6588
   Detection Prevalence : 0.7341
      Balanced Accuracy : 0.7275

       'Positive' Class : 0
```

3. DECISION TREE

A decision tree model is a simple method that can be used to classify objects according to their features.

STEPS

1: First, We draw the decision tree diagram.

2: Look at the confusion matrix.

3: Find the accuracy of the model

4: Plot the ROC curve(receiver operating characteristic curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR))

The accuracy of this model is 0.785.

confusionmatrix	0	1
No	1324	329
Yes	225	232

4. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed.

It can be applied only on numerical variables.

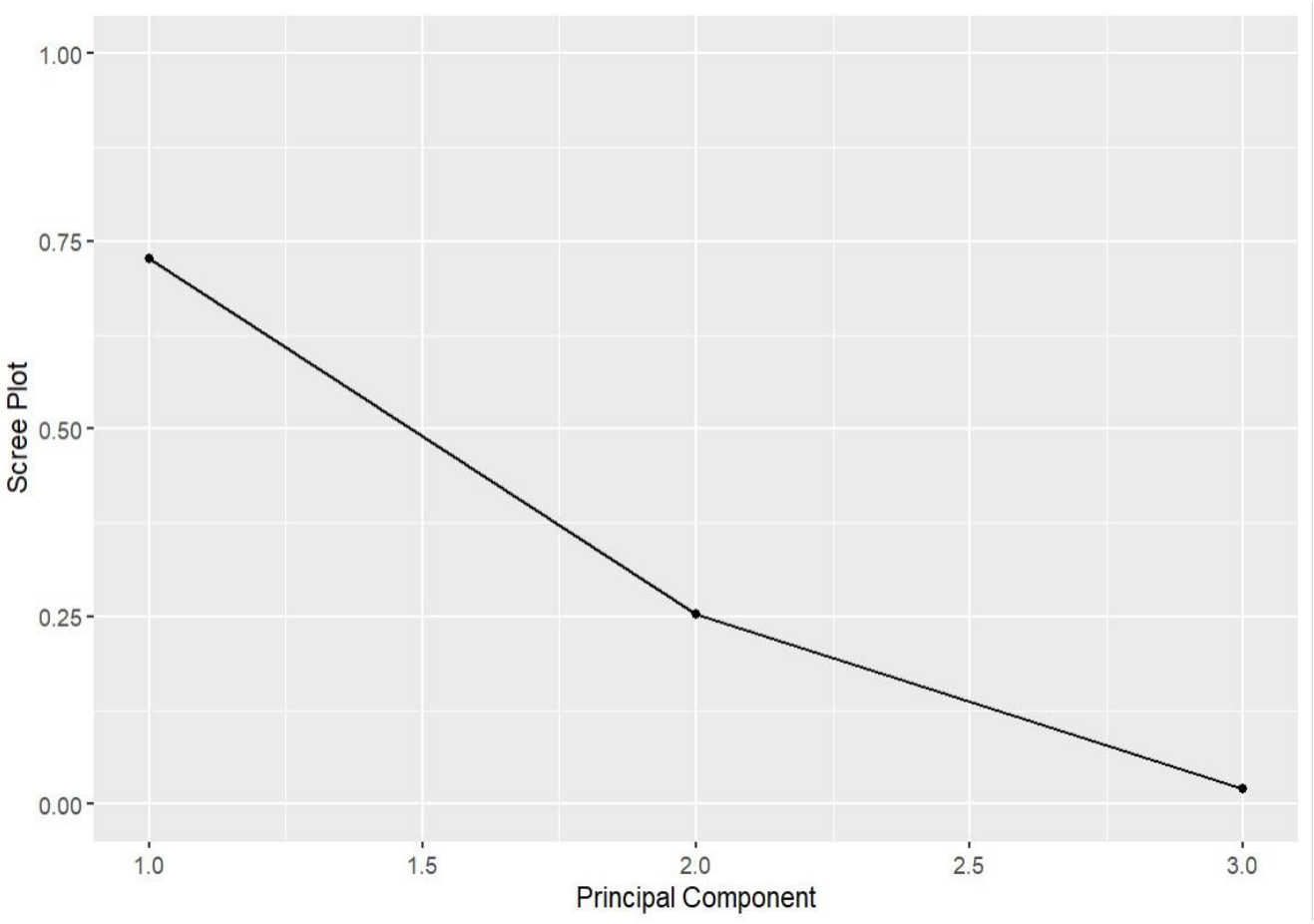
The first principal component has high variation, which indicates that it is significant and the Second Principal component shows high value for Monthly charges.

1. First principal component explains 72.66% of the variation
2. Second principal component explains 25.36% of the variation

3. Third principal component explains 1.98% of the variation.

Here, only the first two explain the majority of the variance in the data.

Finally we plot a scree plot to display the total variance explained by the principal component analysis.



IX. RESULT

All the models that we perform have almost equal accuracy. The SVM produces the low accuracy and Logistic Regression produces the high accuracy.

The below table shows the accuracy obtained by different models.

LOGISTIC REGRESSION	SVM	DECISION TREE
0.869	0.765	0.785

X. Conclusion:

After looking at the model's accuracy measures, logistic regression seems to be the best model when it comes to predicting the number of customers that are likely to churn from telecom industries and hence the companies that would want to track their customer churn, can use this model to receive precise results.

****